November 30, 2023

Classifying English Premier League Match Outcomes: A Model Comparison with COVID-19 Insights

COLLEGE OF CHARLESTON – STATISITICAL LEARNING

ETHAN LEWIS

Table of Contents

Abstract

The following report details the process of fitting, selecting, and comparing a variety of models aimed at predicting the probability of an English Premier League home team victory for a given match using historical match data spanning the 2000/01 to 2022/23 seasons.

In addition to creating test sets for each individual season to track our models' accuracy over the English Premier League's history, we also constructed a COVID-19 test set which contains all matches played 'behind closed doors' from June 20, 2020 to May 17, 2021 to see if our models perform worse when matches are essentially played at a neutral site without 'home field advantage'.

A comprehensive list of all 22 variable abbreviations and definitions can be found on page 14.

Data Source and Cleaning

Original, individual season datasets were sourced from the British sports betting site Football-Data.co.uk and required the following pre-processing steps…

- Drop all betting statistic columns.
    - We are only interested in match data.
- Drop all columns not contained in every individual data set.
    - Older seasons do not track some of the more advanced stats found in recent seasons.
- Convert all *Referee* string values to just the referee's last name.
    - Several formatting inconsistencies across data sets (i.e. first and last name, initials and last name, etc.).
- Convert all *Date* values to YYYY-MM-DD format.
- Create *Day* column using the *Date* column values.

The cleaned, individual season datasets were named and exported as s_YY_YY where the YY placeholders represent the calendar years of each season.

Training and Test Sets (Validation Set Approach)

To create our COVID test set, s_COVID, we iterated through s_19_20 and s_20_21 to remove and store all rows with *Date* values between [2020-06-20, 2021-05-17].

To create the individual season test sets, we randomly sampled 100 rows from each season dataset, naming and storing each result as s_YY_YY_test. The remaining rows not selected to be a member of a test set were combined into a single training set, total_train. Note, we did not apply this process to s_20_21 as it was left with a very small number of rows after creating s_COVID.

Exploratory Data Analysis Part 1

Using **Table 1** which represents the distribution of *FTR* in total_train, we can see the probability of a home team win in the training set is roughly 46.43%.

Table 1	
FTR	**Count**
Home Team Win (1)	2787
Away Team Win or Draw (0)	3215
TOTAL	6002

Next, the scatterplots between all of our numeric predictors (colored by *FTR*) seen in **Figure 1** indicate several strong class separations. As an example, we can notice a distinct separation in the *HST* vs. *AST* scatterplot; intuitively, if a home team records a large number of shots on target it would seem more likely for the home team to win, and vice versa for the away team.

Key
- *FTR* = 1
- *FTR* = 0

Figure 1

Exploratory Data Analysis Part 2

Given the nature of our response variable, it is crucial we choose appropriate types of models capable of answering classification problems. In this report, we will consider the following models:

- Logistic Regression
- Linear Discriminant Analysis (LDA)
- Quadratic Discriminant Analysis (QDA)
- K-Nearest Neighbors (where K = 1, 3, and 5)

The remainder of this section will focus on EDA as it relates to the preparation for fitting our Logistic Regression model. Specifically, we will implement 10-fold cross validation on each predictor to determine if any polynomial transformations (or their inclusion in the model in general) are necessary. This manual variable selection may also dictate the test and training sets we feed our KNN models (discussed further in **Modeling** and **Predicting** sections).

As for the other model types, we will feed the LDA and QDA algorithms the full set of predictors as they attempt to model slight variations of the relationship between our response and predictors and, in LDA's case, perform variable selection.

Below, **Tables 2 and 3** contain the 10-fold cross validation prediction accuracy rates of each predictor when isolated against *FTR*. Numeric predictors (**Table 2**) were each tested with 1st through 4th degree polynomial transformations. **Table 3** is dedicated to the categorical predictors.

Table 2				
Predictor	**1st Degree**	**2nd Degree**	**3rd Degree**	**4th Degree**
HTHG	0.7011024	0.7010993	0.7010952	0.7010892
HTAG	0.6541268	0.6541092	0.6541118	0.6541105
HS	0.5914623	0.5929619	0.5894707	0.5931257
AS	0.5921320	0.5881385	0.5871481	0.5884738
HST	0.6309540	0.6304592	0.6292937	0.6292977
AST	0.6109635	0.6109655	0.6109604	0.6109588
HC	0.5338225	0.5356548	0.5356548	0.5356548
AC	0.5356548	0.5324901	0.5343217	0.5329848
HF	0.5361531	0.5373184	0.5379898	0.5353231
AF	0.5356548	0.5356548	0.5356550	0.5346564
HY	0.5648137	0.5648204	0.5648070	0.5651362
AY	0.5356548	0.5356548	0.5356548	0.5356548
HR	0.5356548	0.5356548	NA	NA
AR	0.5528156	0.5528171	NA	NA

Table 3	
Predictor	**10-Fold CV Accuracy Rate**
Day	0.5363214
HomeTeam	0.6216287
AwayTeam	0.5804751
HTR	0.7539208
Referee	0.5169959

To properly interpret **Tables 2 and 3,** we first need to choose an accuracy rate cutoff we deem meaningful. In our case, a 50% accuracy rate is no different from randomly guessing the match outcome. Thus, we have settled on a 55% accuracy rate cutoff to ensure we are including impactful variables (and to afford us the opportunity to drop some variables from the full model). Selected variables are shaded green, while dropped variables are shaded grey.

Additionally, we can interpret the selected variables in **Table 2** further by noticing how far the green shading extends across the polynomial degree columns. None of our numeric variables perform better when transformed beyond the 1st degree; the accuracy rates remain extremely consistent across transformations.

Modeling

Having completed a thorough exploratory data analysis, we are ready to begin fitting the following models…

1. Logistic Regression (Full Model)

To begin, we will fit a logistic regression model containing all of the predictors listed in **Tables 2 and 3**. This will not only serve as a good baseline model as we progress into dimension reduction, but the full model could ultimately prove valuable considering all of the predictors are above a 50% 10-fold cross validation accuracy rate (even though not all of them met our 55% cutoff).

- **Number of Predictors:** 18
- **AIC:** 5550.6
- **BIC:** 6749.91

2. Logistic Regression (10-Fold Cross-Validation Subset)

Next, we will fit a logistic regression model using the subset of variables and their corresponding transformations we previously identified in **Tables 2 and 3**.

- **Number of Predictors:** 11
- **AIC:** 5583.2
- **BIC:** 6246.44

3. Logistic Regression (Stepwise Selection Subset on AIC)

Now that we have fit our two foundational logistic regression models (**Models 1 and 2**), we can run a forward, backward, and hybrid stepwise selection algorithm on both models to see if we can optimize the AIC even further.

Model 1 (Full Model) ultimately achieved the lowest AIC value during the stepwise selection process and its metrics are listed below. The metrics of the optimal **Model 2** based on AIC are also listed in parentheses for a quick comparison.

- **Number of Predictors:** 12 (8)
- **AIC:** 5473.2 (5575.82)
- **BIC:** 6143.14 (5924.21)

4. Logistic Regression (Stepwise Selection Subset on BIC)

The development of **Model 4** is quite similar to **Model 3**, but now we will use BIC as our decision criteria, rather than AIC.

Again, **Model 1 (Full Model)** ultimately achieved the lowest BIC value during the stepwise selection process and its metrics are listed below. The metrics of the optimal **Model 2** based on BIC are also listed in parentheses for a quick comparison.

- **Number of Predictors:** 8 (5)
- **BIC:** 5571.52 (5628.54)
- **AIC:** 5511.2 (5588.3)

5. K-Nearest Neighbors (where K = 1, 3, and 5)

Having fit and compared **Models 1 – 4**, we can begin to develop a better sense of which set of predictors to feed the KNN algorithm. Our main goal is prediction, so **Model 3**'s large set of predictors seem appealing due to the model's superior AIC. On the other hand, the additional predictors of **Models 1 and 2** may enhance their prediction accuracy. A better decision can be made after seeing the previous 4 model's performances on the test sets in the upcoming, **Predicting**, section.

6. Linear Discriminant Analysis

Next, we will conduct Linear Discriminant Analysis (LDA) using the full set of predictors listed in **Tables 2 and 3** since LDA has the capability to perform dimension reduction on its own.

7. Quadratic Discriminant Analysis

Finally, we will conduct Quadratic Discriminant Analysis (QDA) using the full set of predictors listed in **Tables 2 and 3** as a way to directly compare to **Model 6**.

Predicting

After fitting each model using total_train, we can now begin to understand their Test Classification Error Rates (Test CER) by applying them to each individual s_YY_YY_test, as well as s_COVID.

Since we are dealing with multiple test sets and want to compare models, it would be best to consolidate each model's Test CER results as much as possible. With this in mind, each model listed below reports the following…

- **Figure –** plots individual test set Test CER values.
    - **Average Season Test CER** denoted by dashed blue line.
    - **COVID Test CER** denoted by red dot.
- **Avg. Season Test CER –** model produces a Test CER for each individual s_YY_YY_test and these values are averaged.
- **Total Season Confusion Matrix –** model produces an individual confusion matrix for each s_YY_YY_test and these tables have been aggregated instead of reporting 22 individual matrices.
- **COVID Test CER –** Test CER value produced by model when using s_COVID.
- **COVID Confusion Matrix –** confusion matrix produced by model when using s_COVID.

1. Logistic Regression (Full Model)

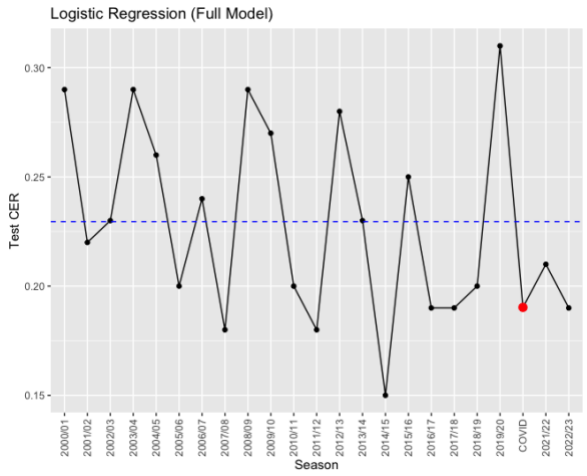

Table 4	
Avg. Season Test CER	0.2295455
Total Seasons Matrix	Predicted *FTR* 0 1 0 943 267 1 238 752
COVID Test CER	0.1902655
COVID Matrix	Predicted *FTR* 0 1 0 232 42 1 44 134

2. Logistic Regression (10-Fold Cross-Validation Subset)

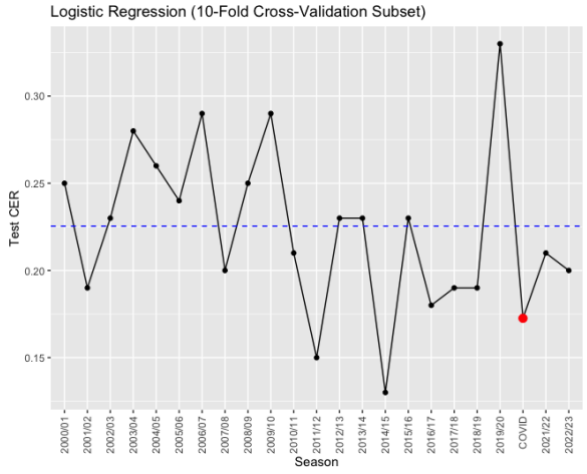

Table 5	
Avg. Season Test CER	0.2254545
Total Seasons Matrix	Predicted *FTR* 0 1 0 955 270 1 226 749
COVID Test CER	0.1725664
COVID Matrix	Predicted *FTR* 0 1 0 240 42 1 36 134

3. Logistic Regression (Stepwise Selection Subset on AIC)

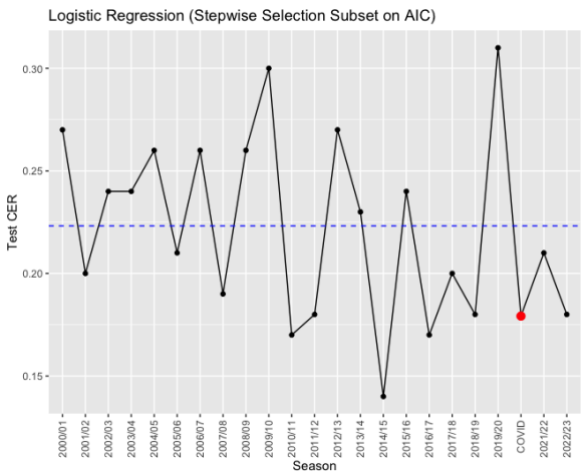

Table 6	
Avg. Season Test CER	0.2295455
Total Seasons Matrix	Predicted *FTR* 0 1 0 949 259 1 232 760
COVID Test CER	0.1792035
COVID Matrix	Predicted *FTR* 0 1 0 238 43 1 38 133

4. Logistic Regression (Stepwise Selection Subset on BIC)

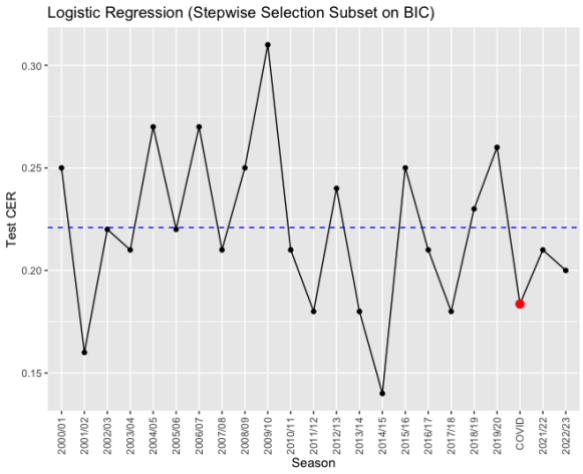

Table 7	
Avg. Season Test CER	0.2209091
Total Seasons Matrix	Predicted *FTR* 0 1 0 960 265 1 221 754
COVID Test CER	0.1836283
COVID Matrix	Predicted *FTR* 0 1 0 239 46 1 37 130

5. K-Nearest Neighbors (where K = 1, 3, and 5)*

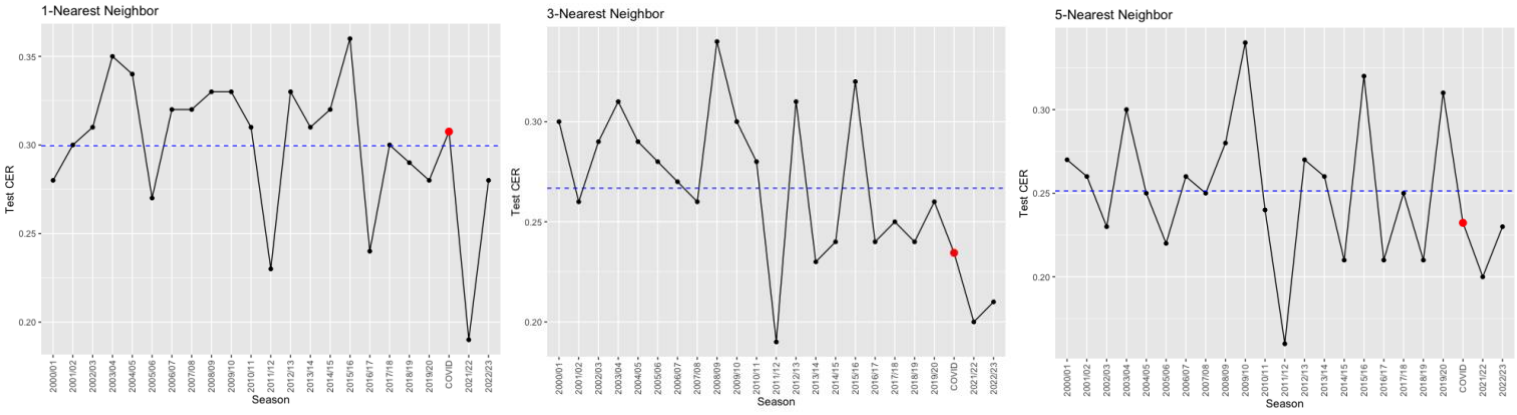

Table 8				
K	**Avg. Season Test CER**	**Total Seasons Matrix**	**COVID Test CER**	**COVID Matrix**
1	0.2995455	Predicted *FTR* 0 1 0 822 335 1 359 684	0.3075221	Predicted *FTR* 0 1 0 193 56 1 83 120
3	0.2668182	0 1 0 877 312 1 304 707	0.2345133	Predicted *FTR* 0 1 0 214 44 1 62 132
5	0.2513636	Predicted *FTR* 0 1 0 907 290 1 274 729	0.2323009	Predicted *FTR* 0 1 0 216 45 1 60 131

*We ultimately chose to fit the KNN models using the 10-fold cross validation (**Model 2**) set of predictors, given that model produced the lowest combination of Season + COVID Test CER values as demonstrated previously in this section.

6. Linear Discriminant Analysis

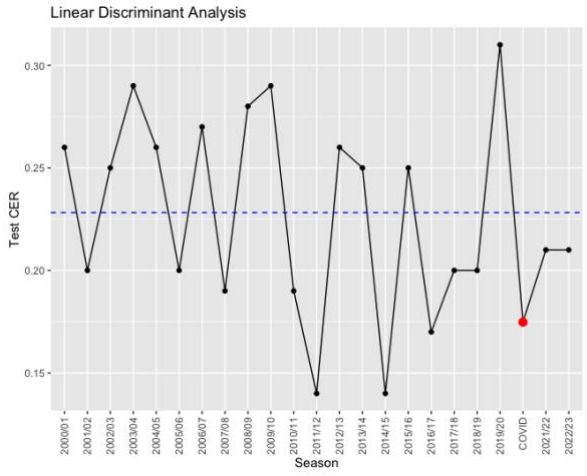

Table 9	
Avg. Season Test CER	0.2281818
Total Seasons Matrix	Predicted *FTR* 0 1 0 968 284 1 212 735
COVID Test CER	0.1747788
COVID Matrix	Predicted *FTR* 0 1 0 241 44 1 35 132

7. Quadratic Discriminant Analysis

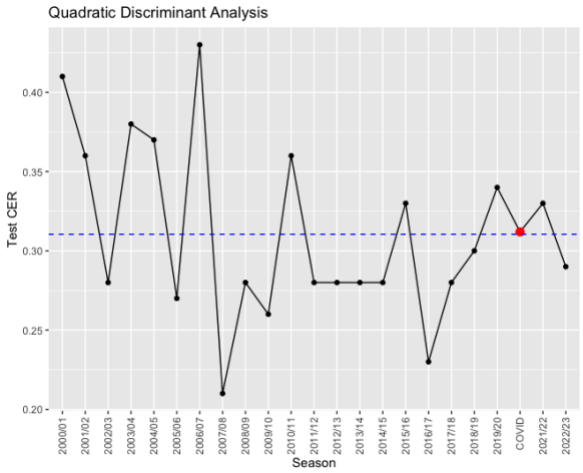

Table 10	
Avg. Season Test CER	0.3104545
Total Seasons Matrix	Predicted *FTR* 0 1 0 815 317 1 366 702
COVID Test CER	0.3119469
COVID Matrix	Predicted *FTR* 0 1 0 210 75 1 66 101

Conclusions

Having applied each of our models to test sets, we can utilize **Tables 11 and 12** below to list each model and their Average Season and COVID Test CER values in increasing order, respectively.

Table 11	
Model	**Average Season Test CER**
Logistic Regression (Stepwise Selection Subset on BIC)	0.2209091
Logistic Regression (Stepwise Selection Subset on AIC)	0.2231818
Logistic Regression (10-Fold Cross Validation Subset)	0.2254545
Linear Discriminant Analysis	0.2281818
Logistic Regression (Full Model)	0.2295455
5-Nearest Neighbors	0.2513636
3-Nearest Neighbors	0.2668182
1-Nearest Neighbor	0.2995455
Quadratic Discriminant Analysis	0.3104545

Table 12	
Model	**COVID Test CER**
Logistic Regression (10-Fold Cross Validation Subset)	0.1725664
Linear Discriminant Analysis	0.1747788
Logistic Regression (Stepwise Selection Subset on AIC)	0.1792035
Logistic Regression (Stepwise Selection Subset on BIC)	0.1836283
Logistic Regression (Full Model)	0.1902655
5-Nearest Neighbors	0.2323009
3-Nearest Neighbors	0.2345133
1-Nearest Neighbor	0.3075221
Quadratic Discriminant Analysis	0.3119469

Both **Tables 11 and 12** indicate two distinct groups of models in terms of Test CER values, which have been delineated with a red line. In **Table 11**, the shift from group 1 to group 2 is represented by a ~2% increase in Test CER, while this same shift in **Table 12** is represented by a ~4% increase in Test CER.

While it is certainly clear the KNN and QDA models are perhaps not the strongest choices for our data, another insight in regard to the Full Logistic Regression model can be made. Although it is a member of the superior group in both **Tables 11 and 12**, the Full Logistic Regression model has consistently performed worse than the other Logistic Regression models only containing a subset of the full range of predictors, as well as LDA which has the capacity to perform variable selection. This would indicate some of the predictors are not extremely impactful and may be distorting the results, like we discussed in the previous **Exploratory Data Analysis Part 2** section.

Beyond comparing the prediction strength among a variety of classification models, we also had the additional question of if our models would perform worse on a COVID test set covering matches from June 20, 2020 to May 17, 2021 due to the lack of fan presence in the home stadiums.

Utilizing the figures in the **Modeling** section, however, we can clearly notice all of our models (apart from 1-NN and QDA) actually produce a better than average Test CER on the COVID test set. If these phenomena weren't surprising enough, it can also be seen the COVID test set consistently has one of the best Test CERs out of every individual season test set for these same models.

It would seem these models are suggesting a surprising conclusion; English Premier League clubs performed just as well, if not better, behind closed doors than in front of their own home crowd.

Variable Dictionary

Response

- ***FTR*: Full Time Result**
    - Home Team Win = 1
    - Away Team Win or Draw = 0

Numeric

- ***HTHG:*** Half Time Home Team Goals
- ***HTAG:*** Half Time Away Team Goals
- ***HS:*** Home Team Shots
- ***AS:*** Away Team Shots
- ***HST:*** Home Team Shots on Target
- ***AST:*** Away Team Shots on Target
- ***HC:*** Home Team Corner Kicks
- ***AC:*** Away Team Corner Kicks
- ***HF:*** Home Team Fouls Committed
- ***AF:*** Away Team Fouls Committed
- ***HY:*** Home Team Yellow Cards Received
- ***AY:*** Away Team Yellow Cards Received
- ***HR:*** Home Team Red Cards Received
- ***AR:*** Away Team Red Cards Received

Categorical

- ***Day:*** Match Day of the Week
- ***HomeTeam:*** Home Team Name
- ***AwayTeam:*** Away Team Name
- ***HTHR:*** Half Time Result
    - Home Team Winning = H
    - Away Team Winning = A
    - Drawing = D
- ***Referee:*** Match Referee Last Name

Other

- ***Date:*** Match Date (YYYY-MM-DD)